F_{ST} and kinship for arbitrary population structures I: Generalized definitions

Alejandro Ochoa[1,2] and John D. Storey[3,*]

[1]Duke Center for Statistical Genetics and Genomics, and [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

∗ Corresponding author: jstorey@princeton.edu

Abstract: F_{ST} is a fundamental measure of genetic differentiation and population structure, currently defined for subdivided populations. F_{ST} in practice typically assumes *independent, non-overlapping subpopulations*, which all split simultaneously from their last common ancestral population so that genetic drift in each subpopulation is probabilistically independent of the other subpopulations. We introduce a generalized F_{ST} definition for arbitrary population structures, where individuals may be related in arbitrary ways, allowing for arbitrary probabilistic dependence among individuals. Our definitions are built on identity-by-descent (IBD) probabilities that relate individuals through inbreeding and kinship coefficients. We generalize F_{ST} as the mean inbreeding coefficient of the individuals' local populations relative to their last common ancestral population. We show that the generalized definition agrees with Wright's original and the independent subpopulation definitions as special cases. We define a novel coancestry model based on "individual-specific allele frequencies" and prove that its parameters correspond to probabilistic kinship coefficients. Lastly, we extend the Pritchard-Stephens-Donnelly admixture model in the context of our coancestry model and calculate its F_{ST}. To motivate this work, we include a summary of analyses we have carried out in follow-up papers, where our new approach has been applied to simulations and global human data, showcasing the complexity of human population structure, demonstrating our success in estimating kinship and F_{ST}, and the shortcomings of existing approaches. The probabilistic framework we introduce here provides a theoretical foundation that extends F_{ST} in terms of inbreeding and kinship coefficients to arbitrary population structures, paving the way for new estimators and novel analyses.

Note: This article is Part I of two-part manuscripts. We refer to these in the text as Part I and Part II, respectively.

Part I: Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (10.1101/083915) (2019). https://doi.org/10.1101/083915. First published 2016-10-27.

Part II: Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". *bioRxiv* (10.1101/083923) (2019). https://doi.org/10.1101/083923. First published 2016-10-27.

Contents

1 Introduction

A population of mating organisms is *structured* if its individuals do not mate randomly, which results in an increase in mean homozygozity over the population compared to that of a randomly mating population [3, 4]. F_{ST} is a parameter that measures population structure [5, 6], which is typically understood through homozygosity. An unstructured population has $F_{\mathrm{ST}} = 0$ and genotypes at each locus have Hardy-Weinberg proportions. At the other extreme, a fully differentiated population has $F_{\mathrm{ST}} = 1$ and every subpopulation at every locus is homozygous for some allele. In addition to measuring population differentiation, F_{ST} is also used to model DNA profile matching uncertainty in forensics [7–13] and to identify loci under selection [14–21]. Current F_{ST} definitions assume a partitioned or subdivided population into discrete, non-overlapping subpopulations [5, 6, 22–24]. Many F_{ST} estimators further assume that subpopulations have evolved independently from the most recent common ancestor (MRCA) population [21–24], which occurs only if every subpopulation split from the MRCA population at the same time (Fig. 1A, Fig. 2A). However, populations such as humans are not naturally subdivided [11, 25–27] (Fig. 1B); thus, arbitrarily imposed subdivisions may yield correlated subpopulations that no longer satisfy the independent subpopulations model assumed by existing F_{ST} estimators (Fig. 2B). In this work, we build a generalized F_{ST} definition applicable to arbitrary population structures, including arbitrary evolutionary dependencies.

Natural populations are often structured due to population size differences and the constraints of distance and geography [31]. For example, the genetic population structure of humans shows evidence of population bottlenecks migrating out of Africa [32–40] as well as numerous admixture events [41–45]. Notably, human populations display genetic similarity that decays smoothly with geographic distance, rather than taking on discrete values as would be expected for independent subpopulations [11, 27, 35, 37–39] (Fig. 1B). Current F_{ST} definitions do not apply to these complex population structures.

F_{ST} is known by many names, including fixation index [6] and coancestry coefficient [23, 46]). F_{ST} is also alternatively defined in terms of the variance of subpopulation allele frequencies [6], variance components [47], correlations [22], and genetic distance [46]. Our generalized F_{ST} is defined using inbreeding coefficients, like Wright's F_{ST}. There is also a diversity of summary statistics that measure locus-specific differentiation, such as G_{ST}, G'_{ST}, and D, which are functions of observed allele frequencies, and which approximate F_{ST} under certain conditions [48–53]. We consider F_{ST} as a genome-wide evolutionary parameter given by relatedness, which modulates the random drift of allele frequencies across loci but does not depend on these frequencies, mutation rates, or other locus-specific features. We review these previous F_{ST} definitions in greater detail in Supplementary Information, Section S1. The focus of our work is to generalize and accurately estimate the genome-wide F_{ST} in individuals with arbitrary relatedness, and does not presently concern locus-specific F_{ST} estimation or the identification of loci under selection.

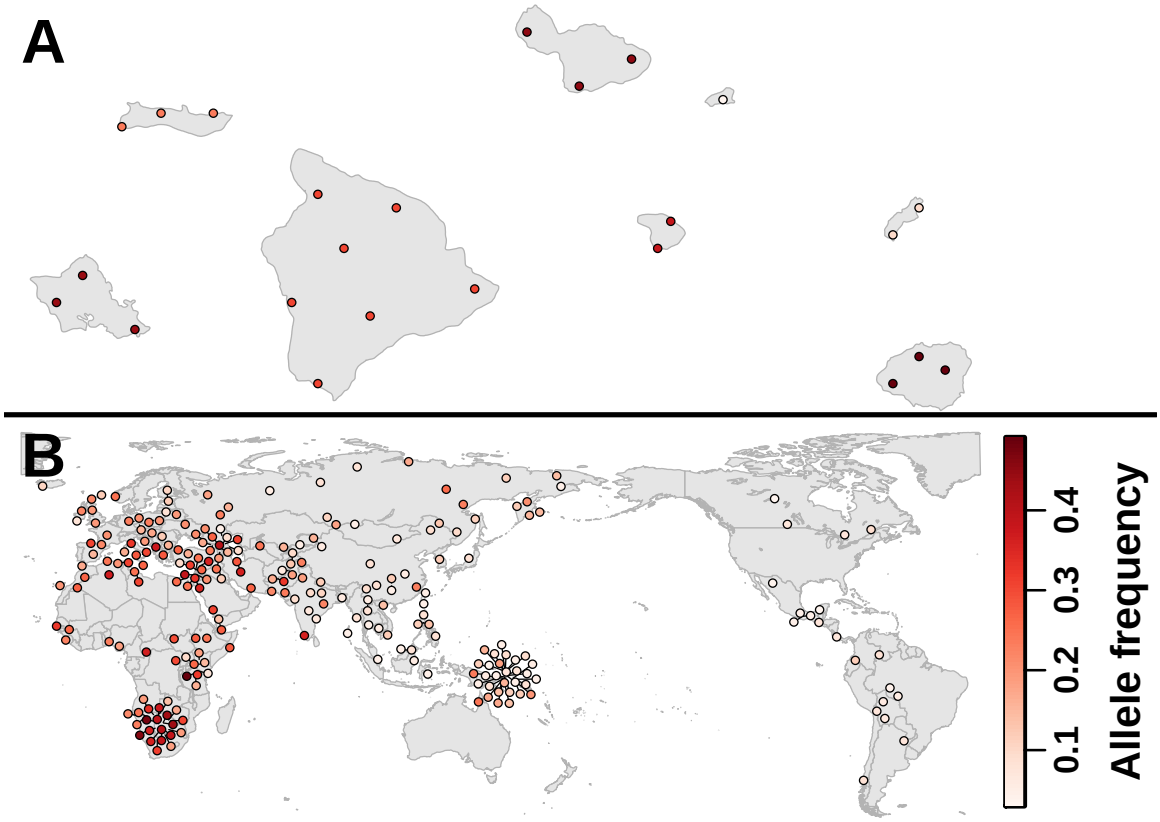

Figure 1: **Independent subpopulations model versus median-F_{ST} human SNP.** In these maps, circles are subpopulations (moved to prevent overlaps in panel B) colored by their mean allele frequency (AF) at a locus. **A.** A simulated locus from the independent subpopulations model illustrated using islands. Individuals from the same island draw their alleles from the same pool, so they have the same underlying AF, while individuals from different islands evolve independently (AFs across islands are uncorrelated). **B.** AFs at SNP locus rs2650044 in the Human Origins datasets of [28–30], illustrates typical differentiation in humans. This locus had the median per-locus F_{ST} estimate (≈ 0.0961) among loci with minor allele frequency $\geq 10\%$ using the estimator of [22] and the $K = 244$ subpopulations shown. Since AFs display strong geographical correlation, the human population does not fit the independent subpopulations model. Subpopulation AFs are estimated using Empirical Bayes (see Supplementary Information, Section S3).

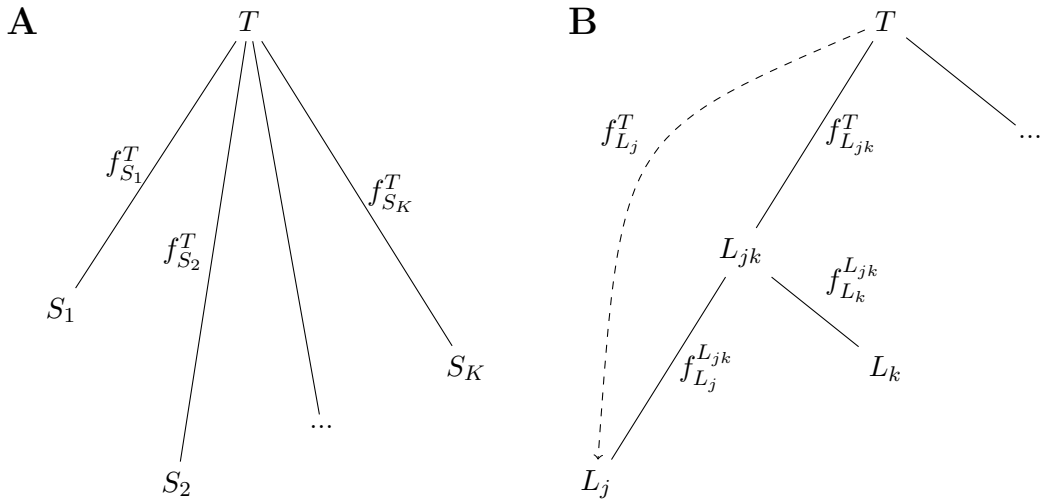

Figure 2: **Independent subpopulations versus arbitrary population structures.** These trees illustrate relationships (edges) between populations (nodes). The edge length between two populations T and S is proportional to the inbreeding coefficient f^T_S (see Section 3.1). **A.** In the independent subpopulations model, K subpopulations $S_1, ..., S_K$ evolved independently from T, which requires that every S_u split from T at the same time. **B.** In an arbitrary population structure, each individual j has its own local population L_j, and every pair of individuals (j, k) have a jointly local population L_{jk} from which L_j and L_k evolved (see Section 3.2). We do not assume a bifurcating tree process: the case for three or more individuals is not a tree (only two individuals j and k are shown). $f^T_{L_j}$ and $f^T_{L_{jk}}$ are relative to T, while $f^{L_{jk}}_{L_j}$, $f^{L_{jk}}_{L_k}$ are relative to L_{jk}.

The developments in this paper have lead to improved estimates of F_{ST} and kinship in Part II [2]. We have also applied these new probabilistic quantities and estimators to data from the Human Origins and 1000 Genomes Project data sets in ref. [54]. To motivate the generalized definitions we present in this work, in Section 2 we provide an overview of simulation results demonstrating the accuracy of the estimators (from Part II) and findings from analyzing the Human Origins and 1000 Genomes Project datasets (from ref [54]). These results establish that a generalized definition of F_{ST} in terms of kinship and inbreeding for arbitrary population structures is needed.

In Section 3 we formally define kinship and inbreeding coefficients, which measure how individuals are related, quantify population structure, and are the foundation of our work. We then generalize F_{ST} in terms of individual parameters (namely, inbreeding coefficients), and in analogy to Wright's F_{IS}, model local inbreeding on an individual basis. Our F_{ST} applies to arbitrary population structures, generalizing previous F_{ST} definitions restricted to subdivided populations.

In Section 4 we show a connection between the coalescent and kinship, inbreeding and generalized F_{ST}. This provides a generalization of a previous result showing the relationship between the coalescent and the classic F_{ST} defined on subdivided populations. In Section 5 we define a coancestry model that parametrizes the correlations of "individual-specific allele frequencies" (IAFs), a recent tool that also accommodates individual-specific relationships [55, 56]. Our model is related to previous models between populations [23, 57]. We prove that our coancestry parameters correspond to kinship coefficients, thereby preserving their probabilistic interpretations, and we relate these parameters to F_{ST}.

Lastly, in Section 6 we provide a novel F_{ST} analysis for admixed individuals by applying our coancestry model from Section 5 to the widely-used Pritchard-Stephens-Donnelly (PSD) admixture model, in which individuals derive their ancestry from several ancestral subpopulations with individual-specific admixture proportions [58–60]. We analyze an extension of the PSD model [55, 61–64] that generates allele frequencies from the Balding-Nichols distribution [7], and propose a more complete coancestry model for the ancestral subpopulations. We derive equations relating F_{ST} to the model parameters of PSD and its extensions. These results enable us to use an admixture simulation without independent subpopulations to benchmark kinship and F_{ST} estimators in Section 6 of Part II.

Our generalized definitions permit the analysis of F_{ST} and kinship estimators under arbitrary population structures, and pave the way forward to new estimation approaches, which are the focus of our following work in this series (Part II).

2 Motivating analyses

The results presented here lead to a deeper understanding of the limitations of existing F_{ST}, kinship, and inbreeding estimators. Specifically, the assumptions underlying existing estimators are too restrictive and do not align with the properties of human populations that have been revealed

through recent studies. In Part II, we theoretically calculate and then numerically verify complex biases that manifest from existing estimators when the population structure and relatedness violates the non-overlapping and independently evolving subpopulations assumptions. This then leads to new estimators of F_{ST}, kinship, and inbreeding proposed in Part II. In ref. [54], we applied the estimators from Part II to data from the Human Origins study and 1000 Genomes Project (TGP). There, it is revealed on these seminal studies that the theory, methods, and simulations from Part I and Part II hold true on real data. Although the results summarized in this section involve details presented in full in Part II and ref. [54], it may be useful to the reader to see the ultimate consequences of the theory present in the current paper, Part I.

In Part II, we carried out simulations in two scenarios. The first scenario approximately satisfies the assumptions of the existing (Weir-Cockerham) estimate of F_{ST}. The second scenario is an admixture model (described in Section 6), which reflects the characteristics we have observed in real data where there are no well-defined independent subpopulations. Fig. 3, columns A and B, show the results of these simulations. It can be seen that that both the existing and proposed estimators do well in the first scenario (Fig. 3A) where the population is divided into non-overlapping subpopulations that have independently evolved from a common ancestral population. However, in the second scenario (Fig. 3B) where these assumptions are violated, the existing estimators show notable downward bias. Our theoretical results determine exactly what this bias is for both kinship and F_{ST}.

In ref. [54], we then analyzed data from the Human Origins [28–30] and TGP studies [65], both of which consist of individuals sampled from a global distribution of ancestries. For the TGP data, we specifically limited our analysis to Hispanics. Our novel kinship estimates calculated on these data reveal a complex population structure in the global human population (Fig. 3C) and in Hispanics in particular (Fig. 3D). Since there are no independent subpopulations in the human data, existing kinship and F_{ST} estimates in these data will also be downwardly biased, which can be seen in the bottom two rows of Fig. 3C-D. In contrast, our more accurate novel F_{ST} estimates measure greater differentiation than has been previously reported (Fig. 3C-D, second and fourth rows). A deeper analysis of our calculations reveals a clear connection between our estimated kinship structure (but not existing estimates) and the global human migrations under the African Origins model [54]. Our results suggest that common population genetic analyses on real human data will greatly benefit from our improved kinship and F_{ST} estimation framework.

3 Generalized definitions in terms of individuals

Now that we have established the need for a more flexible population structure model that does not assume independent subpopulations, we shall introduce here novel definitions required for this goal. First we review the formal definitions of kinship and inbreeding coefficients. Then we define a "local" population for every individual, which allows us to distinguish "structural" inbreeding due to the

population structure from the "local" inbreeding that applies to individuals with closely-related parents. We then introduce our generalized F_{ST} definition as the mean structural inbreeding coefficient, and show that this definition equals the previous F_{ST} definition for independent subpopulations. We also generalize previous formulas for changing the reference ancestral population for kinship and inbreeding coefficients. Lastly, we review the connection between kinship coefficients and the covariance of genotypes.

3.1 Overview of data and model parameters

Table 1 summarizes the notation used in this work. Our models assume that genotypes at every locus evolve neutrally—by random drift only, in the absence of recent mutation and selection. Thus, only the population structure shapes the covariance structure of genotypes.

Let x_{ij} be observed biallelic genotypes for locus $i \in \{1, ..., m\}$ and diploid individual $j \in \{1, ..., n\}$. Given a chosen reference allele at each locus, genotypes are encoded as the number of reference alleles: $x_{ij} = 2$ is homozygous for the reference allele, $x_{ij} = 0$ is homozygous for the alternative allele, and $x_{ij} = 1$ is heterozygous. We focus on biallelic loci since they vastly outnumber other types of genetic variants in humans. Note that a multiallelic model, which would require additional notation, could follow in analogy to previous F_{ST} work for populations [23].

We assume the existence of a panmictic ancestral population T for all individuals under consideration. T is generally not required to be the MRCA population, so many choices of T are possible. Note that T is a collection of organisms ancestral to a given set of individual organisms, shared by all loci, and it is not assumed that the alleles at a given locus coalesce in T. Two alleles are

---

Figure 3 *(following page)*: **New and existing kinship and F_{ST} estimates in simulations and real human data.** Each column corresponds to a given dataset and contains four panels: (1) the true kinship matrix (for simulations only; unknown in real data), (2) our new kinship estimates, (3) the standard kinship estimates, and (4) the comparison of the Weir-Cockerham (WC) F_{ST} estimates to our new F_{ST} estimates and the true F_{ST} value (red dashed lines; unknown in real data). Each kinship matrix plots the kinship values (color scale) between every pair of individuals (x and y axes) and the inbreeding values along the diagonal. **A.** The independent subpopulations simulation is the only scenario where existing F_{ST} estimators are unbiased. The standard kinship estimator has a small bias since the average kinship is fairly low. **B.** A spatial admixture simulation demonstrates biases in existing approaches (distortions in standard kinship estimates and F_{ST} estimates that are half of the true values) and superior performance of our kinship and F_{ST} approach (see Section 6 in Part II for simulation details). **C.** The Human Origins dataset for global human populations. Individuals were grouped into $K = 11$ continental clusters (see [54]). **D.** The Hispanic subset of the 1000 Genomes Project data. Individuals were grouped into $k = 4$ clusters by sampling location (see [54]).

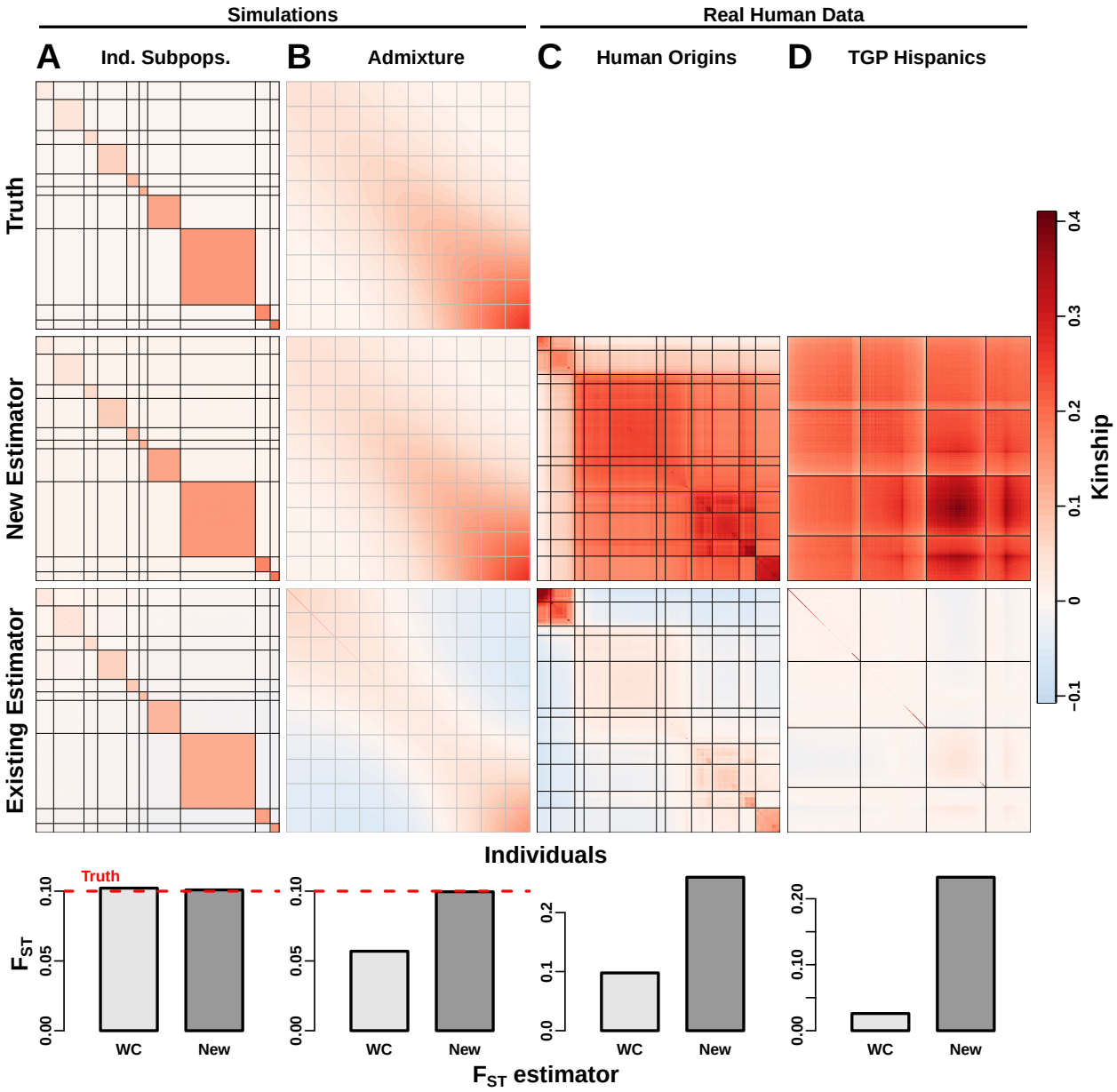

Table 1: **Mathematical notation.**

Symbol	Sec.	Definition
x_{ij}	3.1	Genotype at locus i of individual j, counting the number of reference alleles (0,1,2).
p_i^T	3.1	Reference allele frequency at locus i for population T.
f_j^T	3.1	Inbreeding coefficient: probability that the two alleles at a random locus of individual j are identical by descent (IBD) when the ancestral population is T.
φ_{jk}^T	3.1	Kinship coefficient: probability that two alleles drawn randomly one from individual j and the other from individual k at a random locus are IBD when the ancestral population is T.
f_S^T	3.1	Inbreeding coefficient of the panmictic population S when the ancestral population is T.
L_j	3.2	Local population of individual j.
L_{jk}	3.2	Jointly local population of individuals j and k (MRCA population of L_j and L_k).
$f_j^{L_j}$	3.2	Local inbreeding coefficient of individual j (special case of f_j^T with $T = L_j$).
$\varphi_{jk}^{L_{jk}}$	3.2	Local kinship coefficient of individuals j and k (special case of φ_{jk}^T with $T = L_{jk}$).
$f_{L_j}^T$	3.2	Structural inbreeding coefficient of individual j when the ancestral population is T (special case of f_S^T with $S = L_j$).
$f_{L_{jk}}^T$	3.2	Structural kinship coefficient of individuals j and k when the ancestral population is T (special case of f_S^T with $S = L_{jk}$).
F_{ST}	3.3	Generalized F_{ST}: weighted mean $f_{L_j}^T$ over the individuals in a sample.
$\bar{t}_T$	4	Mean coalescence time for two alleles at a random locus drawn from T.
$\bar{t}_j$	4	Mean coalescence time for the two alleles at a random locus of individual j.
$\bar{t}_{jk}$	4	Mean coalescence time for two alleles drawn at random from each of two individuals j and k.
π_{ij}	5	Individual-specific allele frequency (IAF) at locus i of individual j.
θ_{jk}^T	5	Coancestry coefficient of individuals j and k when the ancestral population is T (equivalent to φ_{jk}^T when $j \neq k$ and to f_j^T when $j = k$).
S_u	6	Supopulation u, or intermediate subpopulation u (when ancestral to admixed individuals).
q_{ju}	6	Admixture proportion of individual j for intermediate subpopulation S_u.
ϑ_{uv}^T	6.2	Coancestry coefficient of the intermediate subpopulations S_u and S_v when the ancestral population is T.

said to be "identical by descent" (IBD) if they originate from a single ancestor organism that lived more recently than the given ancestral population T [4, 6, 66]. In other words, relationships that precede T in time do not count as IBD, while relationships since T count toward IBD probabilities. Every locus i is assumed to have been polymorphic in T, with an ancestral reference allele frequency $p_i^T \in (0,1)$, and no new mutations have occurred since then.

The inbreeding coefficient of individual j relative to T, $f_j^T \in [0,1]$, is defined as the probability that the two alleles of any random locus of j are IBD when the ancestral population is T [67]. Therefore, f_j^T measures the amount of relatedness within an individual, or the extent of dependence between its alleles at each locus. Similarly, the kinship coefficient of individuals j and k relative to T, $\varphi_{jk}^T \in [0,1]$, is defined as the probability that two alleles at any random locus, each picked at random from each of the two individuals, are IBD when the ancestral population is T [5]. φ_{jk}^T measures the amount of relatedness between individuals, or the extent of dependence across their alleles at each locus. Note that children j of parents (k,l) have an expected f_j^T of φ_{kl}^T [5]. Both f_j^T and φ_{jk}^T combine relatedness due to the population structure with recent or "local" relatedness, such as that of family members [68]. The values of f_j^T,φ_{jk}^T are functions of the chosen ancestral population T, which determines the level of relatedness that is treated as unrelated [4, 66]. Thus, f_j^T and φ_{jk}^T increase if T is an earlier rather than a more recent population. The expression "f_j^T relative to T" refers to the value of f_j^T when T is chosen as the reference ancestral population [6, 66]. The mean f_j^T is positive in a structured population [67], and it also increases slowly over time in finite panmictic populations due to genetic drift [69].

Given an ancestral population T (not necessarily the MRCA population in this context) and an unstructured subpopulation S that evolved from T, Malécot defined F_{ST} as the mean f_j^T over the individuals in S relative to T [5], and which we denote by f_S^T. When S is itself structured, Wright defined three coefficients that connect T, S and individuals I in S [6]: F_{IT} ("total inbreeding") is the mean f_j^T of individuals (I) relative to T; F_{IS} ("local inbreeding") is the mean f_j^T of individuals (I) relative to S, which Wright did not consider to be part of the population structure; lastly, F_{ST} ("structural inbreeding") is the mean f_j^T relative to T that would result if individuals in S mated randomly (and which equals our f_S^T). The special case $F_{\mathrm{IS}} = 0$ gives $F_{\mathrm{ST}} = F_{\mathrm{IT}}$ [6]. See Supplementary Information, Section S1.1 for a more detailed review of these definitions. Wright created the distinction between F_{ST} and F_{IT} with animal breeding in mind, since mating systems for artificial selection could cause the local inbreeding (F_{IS}) and therefore also F_{IT} to be large at times, but F_{ST} measures the more relevant mean inbreeding that results after random mating resumes in the strain [67]. However, in large, natural populations F_{IS} is small so $F_{\mathrm{ST}} \approx F_{\mathrm{IT}}$ in these cases. The F_{ST} definition has been extended to a set of disjoint subpopulations, where it is the average F_{ST} of each subpopulation from the last common ancestral population [23, 24].

In practice, the ancestral population T is usually not identified explicitly, which obscures its role in estimating kinship and F_{ST}. Here we clarify this important matter. Every population of mating organisms can be modeled as descending from a panmictic ancestral population T—whether

real or a mathematical construct—that at every locus contained the pool of ancestral alleles that modern individuals inherited. By default, the recommended choice of T is the MRCA population of the individuals in the sample [22–24, 66, 70]. For example, if all individuals are drawn from one effectively panmictic population, then this population is the MRCA. In a pedigree with unrelated founders, the MRCA population consists of these founders [6, 31]. In a population structure defined by a tree, the MRCA population is the root node at which the first split occurs (Fig. 2). The choice of T sets the minimum possible value of φ_{jk}^T: a pair of unrelated individuals drawn from T have $\varphi_{jk}^T = 0$, and an individual from T (with unrelated parents by definition) has $f_j^T = 0$ [71]. Thus, assuming that $\varphi_{jk}^T = 0$ pairs are present in a sample, the set of φ_{jk}^T values is in terms of the MRCA population T if and only if $\min \varphi_{jk}^T = 0$. If $\min \varphi_{jk}^T > 0$, then T is more ancestral than the MRCA population. Estimates with $\min \varphi_{jk}^T < 0$—impossible if φ_{jk}^T is a probability—have an implicit T that is more recent than the MRCA population and cannot be interpreted biologically. For humans, if we ignore the limited Neanderthal and Denisovan introgressions [42, 43], the MRCA population is the real population estimated to have existed in Africa ≈100-200 thousand years ago [32, 33, 40], which first split into the ancestral southern African KhoeSan population (who speak unique "click languages") and the rest of humans [32, 33, 37, 38, 40].

3.2 Local populations

Our generalized F_{ST} definition depends on the notion of a local population. Our formulation includes as special cases the independent subpopulations and admixture models, and its generality is in line with recent efforts to model population structure on a fine scale [72, 73], through continuous spatial models [27, 74–76], or in a manner that makes minimal assumptions [56].

We define the *local population* L_j of an individual j as the MRCA population of j. In the simplest case, if j's parents belong to the same panmictic subpopulation S, then $S = L_j$. However, if j's parents belong to different subpopulations, then L_j is modeled as an admixed population (see example below). More broadly, L_j is the most recent panmictic population from which individual j drew its alleles and its inbreeding coefficient $f_j^{L_j}$ can be meaningfully defined. We define the "local" inbreeding coefficient of j to be $f_j^{L_j}$, and j is said to be *locally outbred* if $f_j^{L_j} = 0$.

For any population T ancestral to L_j, the parameter trio $(f_j^T,\ f_j^{L_j},\ f_{L_j}^T)$ are individual-level analogs of Wright's trio $(F_{\mathrm{IT}},\ F_{\mathrm{IS}},\ F_{\mathrm{ST}})$ defined for a subdivided population [6], with L_j playing the role of S. Moreover, just like Wright's coefficients satisfy

$$(1 - F_{\mathrm{IT}}) = (1 - F_{\mathrm{IS}})\,(1 - F_{\mathrm{ST}})\,, \tag{1}$$

our individual-level parameters satisfy

$$\left(1 - f_j^T\right) = \left(1 - f_j^{L_j}\right)\left(1 - f_{L_j}^T\right), \tag{2}$$

since the probability of the absence of IBD of j relative to T (which is $1 - f_j^T$) equals the product of the independent probabilities of absence of IBD at two levels: of j relative to L_j (which is $1 - f_j^{L_j}$),

and of L_j relative to T (which is $1 - f^T_{L_j}$). Note that an individual j is locally outbred ($f^{L_j}_j = 0$) if and only if $f^T_{L_j} = f^T_j$.

Similarly, we define the *jointly local population* L_{jk} of the pair of individuals j and k as the MRCA population of j and k. Hence, L_{jk} is ancestral to both L_j and L_k (Fig. 2B). We define the "local" kinship coefficient to be $\varphi^{L_{jk}}_{jk}$, and j and k are said to be *locally unrelated* if $\varphi^{L_{jk}}_{jk} = 0$. Since the expected inbreeding coefficient of an individual is the kinship of its parents [5], it follows that locally-unrelated parents have locally-outbred offspring.

Consider an individual j in an admixture model, deriving alleles from two distinct subpopulations A and B with proportions q_{jA} and $q_{jB} = 1 - q_{jA}$. Then L_j is modeled as a population that at locus i has a reference allele frequency of $\pi_{ij} = q_{jA}p^A_i + q_{jB}p^B_i$, where p^A_i and p^B_i are the allele frequencies in A and B, respectively. Considering a pair of individuals (j, k) and varying their admixture proportions, their jointly local population at one extreme is $L_{jk} = L_j = L_k$ if and only if $q_{jA} = q_{kA}$ (in other words, these individuals have the same local population if and only if their admixture proportions are the same); at the other extreme L_{jk} is the MRCA population of A and B if and only if $q_{jA} = 1$ and $q_{kA} = 0$ or vice versa (in other words, these individuals have the most distant jointly local population if and only if they are not admixed and belong to opposite subpopulations).

3.3 The generalized F_{ST} for arbitrary population structures

Recall the individual-level analog of Wright's F_{ST} is $f^T_{L_j}$, which measures the inbreeding coefficient of individual j relative to T due exclusively to the population structure (Fig. 2B, Table 1 and Section 3.2). We generalize F_{ST} for a set of n individuals as

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f^T_{L_j}, \tag{3}$$

where the most meaningful choice of T is the MRCA population of all individuals under consideration, and $w_j > 0, \sum_{j=1}^{n} w_j = 1$ are fixed weights for these individuals. The simplest weights are $w_j = \frac{1}{n}$ for all j. However, we allow for flexibility in the weights so that one may assign them to reflect how individuals were sampled, such as a skewed or uneven sampling scheme. For example, if there are two local populations and the first has twice as many samples as the second, then this can be counteracted by weighing every individual from the first local population half as much as every individual from the second local population. In general, individuals can be weighted inversely proportional to their local population's sample sizes, a scheme used implicitly in the Hudson pairwise F_{ST} estimator [24] and which we iterated for a hierarchy of subdivisions in our analysis of the Human Origins dataset [54]. However, for complex population structures without discrete subpopulations and no obvious sampling biases relative to geography or other variables, we favor uniform weights over complicated weighing schemes (the admixed Hispanic individuals were weighted uniformly in

[54]).

This generalized F_{ST} definition summarizes the population structure with a single value, intuitively measuring the average distance of every individual from T. Moreover, our definition contains the previous F_{ST} definition as a special case, as discussed shortly. For simplicity, we kept Wright's traditional F_{ST} notation rather than using something that resembles our f_S^T notation. A more consistent notation could be $F^T_{\{L_j\}}(\{w_j\})$, which more clearly denotes the weighted average of $f^T_{L_j}$ across individuals. Our definition is more general because the traditional S population is replaced by a set of local populations $\{L_j\}$, which may differ for every individual.

3.3.1 Mean heterozygosity in a structured population

Our generalized F_{ST} parametrizes the reduction in mean heterozygosity relative to the ancestral population T for arbitrary population structures, thus generalizing the familiar connection of the classical F_{ST} to allele fixation in an independently-evolving subpopulation. Here we will assume locally-outbred individuals, for which $f^T_{L_j} = f^T_j$. The expected proportion of heterozygotes H_{ij} of an individual with inbreeding coefficient f^T_j at locus i with an ancestral allele frequency p^T_i is given by [67]

$$H_{ij} = \Pr(x_{ij} = 1|T) = 2p^T_i \left(1 - p^T_i\right)\left(1 - f^T_j\right).$$

The weighted mean of these expected proportion of heterozygotes across individuals, $\bar{H}_i$, is given by our generalized F_{ST}:

$$\bar{H}_i = \sum_j w_j H_{ij} = 2p^T_i \left(1 - p^T_i\right)(1 - F_{\mathrm{ST}}). \tag{4}$$

Hence, individuals have Hardy-Weinberg proportions at every locus ($\bar{H}_i = 2p^T_i \left(1 - p^T_i\right)$) if and only if $F_{\mathrm{ST}} = 0$, which in turn happens if and only if $f^T_j = 0$ for each j. In the other extreme, individuals have fully-fixated alleles at every locus ($\bar{H}_i = 0$), if and only if $F_{\mathrm{ST}} = 1$, which in turn happens if and only if $f^T_j = 1$ for each j.

Eq. (4) presents an apparent paradox since a given sample estimate of the heterozygosity $\bar{H}_i$ on one side does not depend on T, while F_{ST} and p^T_i on the other side vary depending on our choice of ancestral population T. In fact, both sides of Eq. (4) are constant with respect to T under our model: F_{ST} increases as T is taken to be a more distant ancestral population, but p^T_i also changes so that $p^T_i \left(1 - p^T_i\right)(1 - F_{\mathrm{ST}})$ is constant in expectation (see Supplementary Information, Section S4 for a proof of this result).

3.3.2 F_{ST} under the independent subpopulations model

Here we show that our generalized F_{ST} contains as a special case the currently-used F_{ST} definition for independent subpopulations. As discussed above, F_{ST} estimators often assume the independent subpopulations model, in which the population is divided into K non-overlapping subpopulations that evolved independently from their MRCA population T [22–24]. For simplicity, individuals are

often further assumed to be locally outbred and locally unrelated. These assumptions result in the following block structure for our parameters,

$$f_j^T = f_{L_j}^T = f_{S_u}^T \quad \text{for} \quad j \in S_u,$$
$$\varphi_{jk}^T = \begin{cases} f_{S_u}^T & j \in S_u, k \in S_u, j \neq k, \\ 0 & j \in S_u, k \in S_{u'}, u \neq u', \end{cases}$$

where $j, k \in \{1, ..., n\}$ index individuals, $S_u, S_{u'}$ are disjoint subpopulations treated as sets containing individuals, and $u, u' \in \{1, ..., K\}$ index these subpopulations. This population structure corresponds to a tree in which every subpopulation split from T at the same time (Fig. 2A), which is the required demographic scenario that leads to probabilistically-independent subpopulations.

The generalized F_{ST} applied to independent subpopulations agrees with the previous F_{ST} definition of the mean per-subpopulation F_{ST} [23, 24]:

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_{L_j}^T = \sum_{u=1}^{K} \frac{1}{K} f_{S_u}^T,$$

where the weights w_j are such that $\sum_{j \in S_u} w_j = \frac{1}{K}$. Note also that the S_u for $u \in \{1, ..., K\}$ act as the K unique local populations, where $L_j = S_u$ whenever $j \in S_u$.

3.4 IBD probabilities with respect to a reference ancestral population

In developing the generalized F_{ST}, we have made use of equations that relate IBD probabilities in a hierarchy. Here we generalize these equations to individual inbreeding and kinship coefficients, which allow for transformations of these probabilities under a change of reference ancestral population. Our relationships are straightforward generalizations of Wright's equation relating F_{IT}, F_{IS}, and F_{ST} in Eq. (1), now more generally applicable.

Let A be a population ancestral to population B, which is in turn ancestral to population C. The inbreeding coefficients relating every pair of populations in $\{A, B, C\}$ satisfy

$$\left(1 - f_C^A\right) = \left(1 - f_C^B\right)\left(1 - f_B^A\right).$$

A similar form applies for individual inbreeding and kinship coefficients given relative to populations A and B, respectively,

$$\begin{aligned} \left(1 - f_j^A\right) &= \left(1 - f_j^B\right)\left(1 - f_B^A\right), \\ \left(1 - \varphi_{jk}^A\right) &= \left(1 - \varphi_{jk}^B\right)\left(1 - f_B^A\right), \end{aligned} \tag{5}$$

which generalizes Eq. (2). All of these cases follow since the absence of IBD of C (or j, or j, k) relative to A requires independent absence of IBD at two levels: of C (or j, or j, k) relative to B, and of B relative to A. All of the above equations can be extended to a multi-level hierarchy just like Wright did for Eq. (1), by iterating at each level [6].

3.5 Genotype moments under the kinship model

In the kinship model [5, 6, 67, 77], genotypes x_{ij} are random variables with first and second moments given by

$$\mathrm{E}[x_{ij}|T] = 2p_i^T, \tag{6}$$

$$\mathrm{Var}(x_{ij}|T) = 2p_i^T\left(1-p_i^T\right)\left(1+f_j^T\right), \tag{7}$$

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = 4p_i^T\left(1-p_i^T\right)\varphi_{jk}^T. \tag{8}$$

Eq. (6) is a consequence of assuming no selection or new mutations, leaving random drift as the only evolutionary force acting on genotypes [67]. Eq. (7) shows how inbreeding modulates the genotype variance: an outbred individual relative to T ($f_j^T = 0$) has the Binomial variance of $2p_i^T\left(1-p_i^T\right)$ that corresponds to independently-drawn alleles; a fully inbred individual ($f_j^T = 1$) has a scaled Bernoulli variance of $4p_i^T\left(1-p_i^T\right)$ that corresponds to maximally correlated alleles [6]. Lastly, Eq. (8) shows how kinship modulates the correlations between individuals: unrelated individuals relative to T ($\varphi_{jk}^T = 0$) have uncorrelated genotypes, while $\varphi_{jk}^T = 1$ holds for the extreme of identical and fully inbred twins, which have maximally correlated genotypes [5, 77]. Hence, f_j^T and φ_{jk}^T parametrize the frequency of non-independent allele draws within and between individuals. The "self kinship", arising from comparing Eq. (7) to the $j = k$ case in Eq. (8), implies $\varphi_{jj}^T = \frac{1}{2}\left(1+f_j^T\right)$, which is a rescaled inbreeding coefficient resulting from comparing an individual with itself or its identical twin.

4 Kinship and the generalized F_{ST} in terms of the coalescent

Slatkin (1991) [78] derived an expression for the classical F_{ST} (for a subdivided population) in terms of mean coalescence times,

$$F_{\mathrm{ST}} = \frac{\bar{t}_T - \bar{t}_S}{\bar{t}_T},$$

where $\bar{t}_S$ is the mean coalescence time for alleles at a random locus within a subpopulations S, and $\bar{t}_T$ is the mean coalescence time for alleles at a random locus across subpopulations. Here we generalize this expression to encompass inbreeding and kinship coefficients, as well as the generalized F_{ST}.

In all cases that follow, we generalize $\bar{t}_T$ to denote the mean coalescence time for two alleles at a random locus drawn from the ancestral population T; in practice it corresponds to the mean coalescence time of the alleles of the two most distant individuals in the sample. The inbreeding and kinship coefficients are given by

$$f_j^T = \frac{\bar{t}_T - \bar{t}_j}{\bar{t}_T},$$

$$\varphi_{jk}^T = \frac{\bar{t}_T - \bar{t}_{jk}}{\bar{t}_T},$$

where $\bar{t}_j$ is the mean coalescence time of the two alleles of individual j at a random locus, and $\bar{t}_{jk}$ is the mean coalescence time of two alleles drawn at random from each of two individuals j and k at a random locus (see Supplementary Information, Section S2 for derivations). These mean coalescence times could be estimated as average coalescence times for a large number of neutral loci across the genome. If all individuals in the sample are locally outbred, we obtain the desired expression for the generalized F_{ST}:

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_j^T = \frac{\bar{t}_T - \sum_{j=1}^{n} w_j \bar{t}_j}{\bar{t}_T}.$$

Therefore, the generalized F_{ST} equals the relative difference between the weighted mean coalescence times of the alleles within individuals versus the mean coalescence time between the most distantly-related individuals in the sample.

5 The coancestry model for individual allele frequencies

F_{ST} and its estimators are most often studied in terms of subpopulation allele frequencies [22–24, 57]. Here we introduce a coancestry model for individuals, which is based on *individual-specific allele frequencies* (IAFs) [55, 56] that accomodate arbitrary population-level relationships between individuals. Some authors use the terms "coancestry" and "kinship" exchangeably [23, 70, 71]; in our framework, kinship coefficients are general IBD probabilities (following [68]), and we reserve coancestry coefficients for the IAFs covariance parameters (in analogy to the work of [23]). This coancestry model is the foundation behind the extension of the PSD admixture model we present in Section 6 below, and simplifies the analysis of F_{ST} estimator bias in Section 3 of Part II.

In this section we introduce two parameters (see Table 1). First, $\pi_{ij} \in [0, 1]$ is the IAF of individual j at locus i. Individual j draws its two reference alleles independently with probability π_{ij}. Allowing every locus-individual pair to have a potentially-unique allele frequency allows for arbitrary forms of population structure at the level of allele frequencies [56]. Second, $\theta_{jk}^T \in [0, 1]$ is the coancestry coefficient of individuals j and k relative to an ancestral population T, which modulate the covariance of π_{ij} and π_{ik} as shown below.

5.1 The coancestry model

In our coancestry model, the IAFs π_{ij} have the following first and second moments,

$$\mathrm{E}[\pi_{ij}|T] = p_i^T, \tag{9}$$

$$\mathrm{Cov}(\pi_{ij}, \pi_{ik}|T) = p_i^T \left(1 - p_i^T\right) \theta_{jk}^T, \tag{10}$$

$$x_{ij}|\pi_{ij} \sim \mathrm{Binomial}(2, \pi_{ij}). \tag{11}$$

Eq. (9) implies that random drift is the only force acting on the IAFs, and is analogous to Eq. (6) in the kinship model. Eq. (10) is analogous to Eqs. (7) and (8) in the kinship model, with individual

coancestry coefficients (θ^T_{jk}) playing the role of the kinship and inbreeding coefficients (for $j = k$), a relationship elaborated in the next section. Lastly, Eq. (11) draws the two alleles of a genotype independently from the IAF, which models locally-outbred ($f_j^{L_j} = 0$) and locally-unrelated ($\varphi_{jk}^{L_{jk}} = 0$) individuals [23]. Hence, the coancestry model excludes local relationships, so it is more restrictive than the kinship model.

Our coancestry model between individuals is closely related to previous models between subpopulations [23, 57]. However, previous models allowed $\theta^T_{jk} < 0$ [23]. We require that $\theta^T_{jk} \in [0, 1]$ for two reasons: (1) covariance is non-negative in latent structure models [79], such as population structure, and (2) it is necessary in order to relate θ^T_{jk} to IBD probabilities as shown next.

5.2 Relationship between coancestry and kinship coefficients

Here we show that the coancestry coefficients for IAFs, θ_{jk}, defined above can be written in terms of the kinship and inbreeding coefficients utilized in our more general model. We do so by relating our coancestry coefficients to general kinship coefficients by matching moments. Conditional on the IAFs, genotypes in the coancestry model have a Binomial distribution, so

$$\begin{aligned} \mathrm{E}[x_{ij}|\pi_{ij}] &= 2\pi_{ij}, \\ \mathrm{Cov}(x_{ij}, x_{ik}|\pi_{ij}, \pi_{ik}) &= \begin{cases} 2\pi_{ij}(1-\pi_{ij}) & j = k \\ 0 & j \neq k \end{cases}. \end{aligned}$$

We calculate total moments by marginalizing the IAFs. The total expectation is

$$\mathrm{E}[x_{ij}|T] = \mathrm{E}[\mathrm{E}[x_{ij}|\pi_{ij}]|T] = \mathrm{E}[2\pi_{ij}|T] = 2p_i^T,$$

which agrees with Eq. (6) of the kinship model. The total covariance is calculated using

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = \mathrm{E}\left[\mathrm{Cov}(x_{ij}, x_{ik}|\pi_{ij}, \pi_{ik})|T\right] + \mathrm{Cov}\left(\mathrm{E}[x_{ij}|\pi_{ij}], \mathrm{E}[x_{ik}|\pi_{ik}]|T\right).$$

The first term is zero for $j \neq k$, and for $j = k$ it is

$$\begin{aligned} \mathrm{E}\left[\mathrm{Var}(x_{ij}|\pi_{ij})|T\right] &= \mathrm{E}\left[2\pi_{ij}(1-\pi_{ij})|T\right] \\ &= 2\left(\mathrm{E}[\pi_{ij}] - \mathrm{Var}(\pi_{ij}|T) - \mathrm{E}[\pi_{ij}]^2\right) \\ &= 2p_i^T(1-p_i^T)\left(1-\theta^T_{jj}\right) \end{aligned}$$

The second term equals $4\,\mathrm{Cov}\left(\pi_{ij}, \pi_{ik}|T\right)$ for all (j, k) cases, which is given by Eq. (10). All together,

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = \begin{cases} 2p_i^T(1-p_i^T)\left(1+\theta^T_{jj}\right) & j = k, \\ 4p_i^T\left(1-p_i^T\right)\theta^T_{jk} & j \neq k. \end{cases}$$

Comparing the above to Eqs. (7) and (8), we find that

$$\theta^T_{jk} = \begin{cases} f^T_j & \text{if} \quad j = k, \\ \varphi^T_{jk} & \text{if} \quad j \neq k. \end{cases} \tag{12}$$

Therefore, our coancestry coefficients are equal to kinship coefficients, except that self-coancestries are equal to inbreeding coefficients.

Since individuals in our IAF coancestry model are locally outbred and locally unrelated, we also have $f^T_{L_j} = \theta^T_{jj}$ and $f^T_{L_{jk}} = \theta^T_{jk}$ for $j \neq k$. Replacing these quantities in Eq. (3), we obtain the generalized F_{ST} in terms of coancestry coefficients.

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \theta^T_{jj}. \tag{13}$$

6 Coancestry and F_{ST} in admixture models

The Pritchard-Stephens-Donnelly (PSD) admixture model [58] is a well-established, tractable model of structure that is more complex than the independent subpopulations model. There are several algorithms available to estimate the PSD model parameters [58–60, 64, 80]. This model assumes the existence of several intermediate ancestral subpopulations, from which individuals draw alleles according to their admixture proportions. However, the PSD model was not developed with F_{ST} in mind; we will present a modified model that is compatible with our coancestry model. The results presented in this section are applied to evaluate kinship and F_{ST} estimators in Section 6 of Part II, where an admixed population without independent subpopulations is simulated and the true kinship and F_{ST} are known.

The PSD model is a special case of our coancestry model with the following additional parameters (see Table 1). The number of intermediate subpopulations is denoted by K. Let $p_i^{S_u} \in [0, 1]$ be the reference allele frequency at locus i and intermediate subpopulation S_u ($u \in \{1, ..., K\}$; compare $p_i^{S_u}$ to previous notation p_i^T in Table 1). Lastly, $q_{ju} \in [0, 1]$ is the admixture proportion of individual j for intermediate subpopulation S_u. These proportions satisfy $\sum_{u=1}^{K} q_{ju} = 1$ for each j.

6.1 The PSD model with Balding-Nichols allele frequencies

The original algorithm for fitting the PSD model [58] utilizes prior distributions for intermediate subpopulation allele frequencies and admixture proportions according to

$$(q_{ju})_{u=1}^{K} \sim \text{Dirichlet}\,(\alpha, ..., \alpha)\,, \tag{14}$$

$$p_i^{S_u} \sim \text{Beta}(1, 1). \tag{15}$$

Subsequent work has shown [56, 60] that the PSD model of [58] is then equivalent to forming IAFs

$$\pi_{ij} = \sum_{u=1}^{K} p_i^{S_u} q_{ju} \tag{16}$$

where genotypes are then drawn independently according to $x_{ij}|\pi_{ij} \sim \text{Binomial}(2, \pi_{ij})$.

Here we consider an extension of this model, which we call the "BN-PSD" model, by replacing Eq. (15) with the Balding-Nichols (BN) distribution [7] to generate the allele frequencies $p_i^{S_u}$ for the intermediate subpopulations from their MRCA population T. The BN-PSD model establishes an independent subpopulations structure of the intermediate subpopulations S_u as illustrated in Fig. 4. This combined model has been used to simulate structured genotypes [55, 62, 63], and is the target of some inference algorithms [61, 64]. The BN distribution is the following reparametrized Beta distribution,

$$p^* \sim \text{BN}(p, F) = \text{Beta}\left(p\left(\frac{1}{F} - 1\right), (1-p)\left(\frac{1}{F} - 1\right)\right),$$

where p is the ancestral allele frequency and F is the inbreeding coefficient [7]. The resulting allele frequencies p^* fit into our coancestry model, since $\text{E}[p^*] = p$ and $\text{Var}(p^*) = p(1-p)F$.

In BN-PSD, the allele frequencies $p_i^{S_u}$ at each locus i for intermediate subpopulation S_u are drawn independently from

$$p_i^{S_u}|T \sim \text{BN}\left(p_i^T, f_{S_u}^T\right),$$

where p_i^T is the ancestral allele frequency and $f_{S_u}^T$ is the inbreeding coefficient of S_u relative to T (compare $f_{S_u}^T$ to f_S^T notation in Table 1).

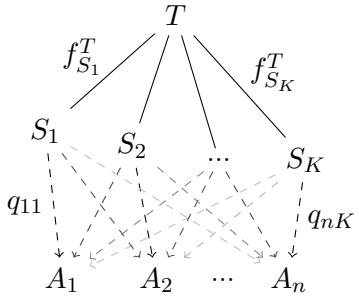

Figure 4: **The demographic model of the BN-PSD admixture model.** There are K intermediate subpopulations S_u for $u \in \{1, ..., K\}$ that evolved independently from their MRCA population T, each of which has its own inbreeding coefficient $f_{S_u}^T$ (solid edges). There are n admixed individuals denoted as A_j for $j \in \{1, ..., n\}$, each deriving ancestry from the intermediate subpopulation S_u (dashed arrows with variable shading) in proportion q_{ju} (i.e. an expected fraction q_{ju} of alleles of individual j are drawn from the intermediate subpopulation S_u).

We calculate the coancestry parameters of this model by matching moments conditional on the admixture proportions $\mathbf{Q} = (q_{ju})$. We calculate the expectation as

$$\mathrm{E}[\pi_{ij}|\mathbf{Q}, T] = \sum_{u=1}^{K} q_{ju}\, \mathrm{E}\left[p_i^{S_u}\middle|T\right] = \sum_{u=1}^{K} q_{ju} p_i^T = p_i^T.$$

and the IAF covariance is

$$\mathrm{Cov}(\pi_{ij}, \pi_{ik}|\mathbf{Q}, T) = \sum_{u=1}^{K} q_{ju} q_{ku}\, \mathrm{Var}\left(p_i^{S_u}\middle|T\right) = p_i^T(1-p_i^T)\sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^T.$$

By matching these to Eq. (10), we arrive at coancestry coefficients and F_{ST} of

$$\begin{aligned} \theta_{jk}^T &= \sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^T, \\ F_{\mathrm{ST}} &= \sum_{j=1}^{n}\sum_{u=1}^{K} w_j q_{ju}^2 f_{S_u}^T. \end{aligned} \tag{17}$$

6.2 The BN-PSD model with full coancestry

The BN-PSD model contains a restriction that the K intermediate subpopulations are independent. Suppose instead that the intermediate subpopulation allele frequencies $p_i^{S_u}$ satisfy our more general coancestry model:

$$\begin{aligned} \mathrm{E}\left[p_i^{S_u}\middle|T\right] &= p_i^T, \\ \mathrm{Cov}\left(p_i^{S_u}, p_i^{S_v}\middle|T\right) &= p_i^T\left(1-p_i^T\right)\vartheta_{uv}^T, \end{aligned}$$

where ϑ_{uv}^T is the coancestry of the intermediate subpopulations S_u and S_v. Note that the previous BN-PSD model satisfies $\vartheta_{uu}^T = f_{S_u}^T$ and $\vartheta_{uv}^T = 0$ for $u \neq v$. Repeating our calculations assuming our full coancestry setting, individual coancestry coefficients and F_{ST} are given by

$$\theta_{jk}^T = \sum_{u=1}^{K}\sum_{v=1}^{K} q_{ju} q_{kv} \vartheta_{uv}^T, \tag{18}$$

$$F_{\mathrm{ST}} = \sum_{j=1}^{n}\sum_{u=1}^{K}\sum_{v=1}^{K} w_j q_{ju} q_{jv} \vartheta_{uv}^T. \tag{19}$$

Therefore, all coancestry coefficients of the intermediate subpopulations influence the individual coancestry coefficients and the overall F_{ST}. The form for θ_{jk}^T above has a simple probabilistic interpretation: the probability of IBD at random loci between individuals j and k corresponds to the sum for each pair of subpopulations u and v of the probability of the pairing $(q_{ju}q_{kv})$ times the probability of IBD between these subpopulations (ϑ_{uv}^T). Note that Eq. (18) was derived independently for a related model [81], but the value of F_{ST} for a set of admixed individuals—which we provide in Eq. (19)—had not been described before to the best of our knowledge.

7 Discussion

We presented a generalized F_{ST} definition corresponding to a weighted mean of individual-specific inbreeding coefficients. Compared to previous F_{ST} definitions, ours is applicable to arbitrary population structures, and in particular does not require the existence of non-overlapping subpopulations.

We considered two closely-related population structure models with individual-level resolution: the kinship model for genotypes, and our new coancestry model for IAFs (individual-specific allele frequencies). The kinship model is the most general, applicable to the genotypes in arbitrary sets of individuals. Our IAF model requires a local form of Hardy-Weinberg equilibrium, and it does not model locally-related or locally-inbred individuals. Nevertheless, IAFs arise in many applications, including admixture models [59], estimation of local kinship [55], genome-wide association studies [82], and the logistic factor analysis [56]. We prove that kinship coefficients, which control genotype covariance, also control IAF covariance under our coancestry model.

We also calculated F_{ST} for admixture models. To achieve this, we framed the PSD (Pritchard-Stephens-Donnelly) admixture model as a special case of our IAF coancestry model, and studied extensions where the intermediate subpopulations are more structured. F_{ST} was previously studied in an admixture model under Nei's F_{ST} definition for one locus, where F_{ST} in the admixed population is given by a ratio involving admixture proportions and intermediate subpopulation allele frequencies [52]. On the other hand, our F_{ST} is an IBD probability shared by all loci and independent of allele frequencies. Under our framework, the F_{ST} of an admixed individual is a sum of products, which is quadratic in the admixture proportions and linear in the coancestry coefficients of the intermediate subpopulations. In the future, inference algorithms for our admixture model with fully-correlated intermediate subpopulations could yield improved results, including coancestry and F_{ST} estimates.

Our probabilistic model reconnects F_{ST} [21, 23, 24] to inbreeding and kinship coefficients [68, 70, 83, 84], all quantities of great interest in population genetics, but which are currently studied in isolation. The main reason for this isolation is that F_{ST} estimation assumes the independent subpopulations model, in which kinship coefficients are uninteresting. However, study of the generalized F_{ST} in arbitrary population structures requires the consideration of arbitrary kinship coefficients [68]. Our work lays the foundation necessary to study estimation of the generalized F_{ST}, which is the focus of our next publication in this series (Part II).

Acknowledgments

This research was supported in part by NIH grant R01 HG006448.

References

[1] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (10.1101/083915) (2019). `https://doi.org/10.1101/083915`. First published 2016-10-27.

[2] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". *bioRxiv* (10.1101/083923) (2019). `https://doi.org/10.1101/083923`. First published 2016-10-27.

[3] Sewall Wright. "Coefficients of Inbreeding and Relationship". *The American Naturalist* 56(645) (1922), pp. 330–338.

[4] Douglas S. Falconer and Trudy F. C. Mackay. *Introduction to Quantitative Genetics*. 4 edition. Harlow: Pearson, 1996. 480 pp.

[5] Gustave Malécot. *Mathématiques de l'hérédité*. Masson et Cie, 1948.

[6] S. Wright. "The genetical structure of populations". *Ann Eugen* 15(4) (1951), pp. 323–354.

[7] D. J. Balding and R. A. Nichols. "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity". *Genetica* 96(1) (1995), pp. 3–12.

[8] David J. Balding. "Likelihood-based inference for genetic correlation coefficients". *Theoretical Population Biology*. Uses of DNA and genetic markers for forensics and population studies 63(3) (2003), pp. 221–230.

[9] Mari Nelis et al. "Genetic Structure of Europeans: A View from the North–East". *PLOS ONE* 4(5) (2009), e5472.

[10] Nuno M. Silva et al. "Human Neutral Genetic Variation and Forensic STR Data". *PLOS ONE* 7(11) (2012), e49666.

[11] Christopher D. Steele, Denise Syndercombe Court, and David J. Balding. "Worldwide FST Estimates Relative to Five Continental-Scale Populations". *Annals of Human Genetics* 78(6) (2014), pp. 468–477.

[12] Bruce Weir and Xiuwen Zheng. "SNPs and SNVs in forensic science". *Forensic Science International: Genetics Supplement Series* 5 (Dec 2015), e267–e268.

[13] John Buckleton et al. "Population-specific FST values for forensic STR markers: A worldwide survey". *Forensic Science International: Genetics* 23 (2016), pp. 91–100.

[14] L. L. Cavalli-Sforza. "Population Structure and Human Evolution". *Proceedings of the Royal Society of London. Series B, Biological Sciences* 164(995) (1966), pp. 362–379.

[15] R. C. Lewontin and Jesse Krakauer. "Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms". *Genetics* 74(1) (1973), pp. 175–195.

[16] Mark A. Beaumont and Richard A. Nichols. "Evaluating Loci for Use in the Genetic Analysis of Population Structure". *Proceedings of the Royal Society of London B: Biological Sciences* 263(1377) (1996), pp. 1619–1626.

[17] Renaud Vitalis, Kevin Dawson, and Pierre Boursot. "Interpretation of Variation Across Marker Loci as Evidence of Selection". *Genetics* 158(4) (2001), pp. 1811–1823.

[18] Joshua M. Akey et al. "Interrogating a High-Density SNP Map for Signatures of Natural Selection". *Genome Res.* 12(12) (2002), pp. 1805–1814.

[19] Adam H. Porter. "A test for deviation from island-model population structure". *Molecular Ecology* 12(4) (2003), pp. 903–915.

[20] Mark A. Beaumont and David J. Balding. "Identifying adaptive genetic divergence among populations from genome scans". *Molecular Ecology* 13(4) (2004), pp. 969–980.

[21] Matthieu Foll and Oscar Gaggiotti. "A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective". *Genetics* 180(2) (2008), pp. 977–993.

[22] B. S. Weir and C. Clark Cockerham. "Estimating F-Statistics for the Analysis of Population Structure". *Evolution* 38(6) (1984), pp. 1358–1370.

[23] B. S. Weir and W. G. Hill. "Estimating F-Statistics". *Annual Review of Genetics* 36(1) (2002), pp. 721–750.

[24] Gaurav Bhatia et al. "Estimating and interpreting FST: the impact of rare variants". *Genome Res.* 23(9) (2013), pp. 1514–1521.

[25] R. C. Lewontin. "The Apportionment of Human Diversity". *Evolutionary Biology*. Ed. by Theodosius Dobzhansky, Max K. Hecht, and William C. Steere. Springer US, 1995, pp. 381–398.

[26] Guido Barbujani et al. "An apportionment of human DNA diversity". *PNAS* 94(9) (1997), pp. 4516–4519.

[27] John Novembre et al. "Genes mirror geography within Europe". *Nature* 456(7218) (2008), pp. 98–101.

[28] Iosif Lazaridis et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". *Nature* 513(7518) (2014), pp. 409–413.

[29] Iosif Lazaridis et al. "Genomic insights into the origin of farming in the ancient Near East". *Nature* 536(7617) (2016), pp. 419–424.

[30] Pontus Skoglund et al. "Genomic insights into the peopling of the Southwest Pacific". *Nature* 538(7626) (2016), pp. 510–513.

[31] Sewall Wright. "Isolation by Distance". *Genetics* 28(2) (1943), pp. 114–138.

[32] Yu-Sheng Chen et al. "mtDNA Variation in the South African Kung and Khwe—and Their Genetic Relationships to Other African Populations". *The American Journal of Human Genetics* 66(4) (2000), pp. 1362–1383.

[33] Lev A. Zhivotovsky, Noah A. Rosenberg, and Marcus W. Feldman. "Features of Evolution and Expansion of Modern Humans, Inferred from Genomewide Microsatellite Markers". *The American Journal of Human Genetics* 72(5) (2003), pp. 1171–1186.

[34] Gabor T Marth et al. "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations." *Genetics* 166(1) (2004), pp. 351–372.

[35] Sohini Ramachandran et al. "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa". *Proc Natl Acad Sci U S A* 102(44) (2005), pp. 15942–15947.

[36] Matthieu Foll and Oscar Gaggiotti. "Identifying the Environmental Factors That Determine the Genetic Structure of Populations". *Genetics* 174(2) (2006), pp. 875–891.

[37] Jun Z. Li et al. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation". *Science* 319(5866) (2008), pp. 1100–1104.

[38] Sarah A. Tishkoff et al. "The Genetic Structure and History of Africans and African Americans". *Science* 324(5930) (2009), pp. 1035–1044.

[39] Graham Coop et al. "The Role of Geography in Human Adaptation". *PLoS Genet* 5(6) (2009), e1000500.

[40] G. David Poznik et al. "Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females". *Science* 341(6145) (2013), pp. 562–565.

[41] Noah A. Rosenberg et al. "Genetic Structure of Human Populations". *Science* 298(5602) (2002), pp. 2381–2385.

[42] Richard E. Green et al. "A draft sequence of the Neandertal genome". *Science* 328(5979) (2010), pp. 710–722.

[43] David Reich et al. "Genetic history of an archaic hominin group from Denisova Cave in Siberia". *Nature* 468(7327) (2010), pp. 1053–1060.

[44] Joseph K. Pickrell and Jonathan K. Pritchard. "Inference of population splits and mixtures from genome-wide allele frequency data". *PLoS Genet.* 8(11) (2012), e1002967.

[45] Nick Patterson et al. "Ancient admixture in human history". *Genetics* 192(3) (2012), pp. 1065–1093.

[46] John Reynolds, B. S. Weir, and C. Clark Cockerham. "Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance". *Genetics* 105(3) (1983), pp. 767–779.

[47] C. Clark Cockerham. "Variance of Gene Frequencies". *Evolution* 23(1) (1969), pp. 72–84.

[48] Masatoshi Nei. "Analysis of Gene Diversity in Subdivided Populations". *PNAS* 70(12) (1973), pp. 3321–3323.

[49] Philip W. Hedrick. "A Standardized Genetic Differentiation Measure". *Evolution* 59(8) (2005), pp. 1633–1638.

[50] Lou Jost. "GST and its relatives do not measure differentiation". *Molecular Ecology* 17(18) (2008), pp. 4015–4026.

[51] Michael C. Whitlock. "Gst' and D do not replace FST". *Molecular Ecology* 20(6) (2011), pp. 1083–1091.

[52] Simina M. Boca and Noah A. Rosenberg. "Mathematical properties of between admixed populations and their parental source populations". *Theoretical Population Biology* 80(3) (2011), pp. 208–216.

[53] Mattias Jakobsson, Michael D. Edge, and Noah A. Rosenberg. "The Relationship Between FST and the Frequency of the Most Frequent Allele". *Genetics* 193(2) (2013), pp. 515–528.

[54] Alejandro Ochoa and John D. Storey. "New kinship and F_{ST} estimates reveal higher levels of differentiation in the global human population". *bioRxiv* (10.1101/653279) (2019). `https://doi.org/10.1101/653279`.

[55] Timothy Thornton et al. "Estimating kinship in admixed populations". *Am. J. Hum. Genet.* 91(1) (2012), pp. 122–138.

[56] Wei Hao, Minsun Song, and John D. Storey. "Probabilistic models of genetic variation in structured populations applied to global human studies". *Bioinformatics* 32(5) (2016), pp. 713–721.

[57] George Nicholson et al. "Assessing population differentiation and isolation from single-nucleotide polymorphism data". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4) (2002), pp. 695–715.

[58] J. K. Pritchard, M. Stephens, and P. Donnelly. "Inference of population structure using multilocus genotype data". *Genetics* 155(2) (2000), pp. 945–959.

[59] Hua Tang et al. "Estimation of individual admixture: analytical and study design considerations". *Genet. Epidemiol.* 28(4) (2005), pp. 289–301.

[60] David H. Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9) (2009), pp. 1655–1664.

[61] Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard. "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". *Genetics* 164(4) (2003), pp. 1567–1587.

[62] Alkes L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". *Nat. Genet.* 38(8) (2006), pp. 904–909.

[63] Timothy Thornton and Mary Sara McPeek. "ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure". *Am. J. Hum. Genet.* 86(2) (2010), pp. 172–184.

[64] Anil Raj, Matthew Stephens, and Jonathan K. Pritchard. "fastSTRUCTURE: variational inference of population structure in large SNP data sets". *Genetics* 197(2) (2014), pp. 573–589.

[65] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing". *Nature* 467(7319) (2010), pp. 1061–1073.

[66] Elizabeth A. Thompson. "Identity by descent: variation in meiosis, across genomes, and in populations". *Genetics* 194(2) (2013), pp. 301–326.

[67] Sewall Wright. "Systems of Mating. IV. the Effects of Selection". *Genetics* 6(2) (1921), pp. 162–166.

[68] William Astle and David J. Balding. "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4) (2009). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

[69] Sewall Wright. "Evolution in Mendelian Populations". *Genetics* 16(2) (1931), pp. 97–159.

[70] Doug Speed and David J. Balding. "Relatedness in the post-genomic era: is it still useful?" *Nat. Rev. Genet.* 16(1) (2015), pp. 33–44.

[71] Bruce S. Weir, Amy D. Anderson, and Amanda B. Hepler. "Genetic relatedness analysis: modern data and new challenges". *Nat Rev Genet* 7(10) (2006), pp. 771–780.

[72] Daniel John Lawson et al. "Inference of population structure using dense haplotype data". *PLoS Genet.* 8(1) (2012), e1002453.

[73] Garrett Hellenthal et al. "A Genetic Atlas of Human Admixture History". *Science* 343(6172) (2014), pp. 747–751.

[74] Gilles Guillot et al. "A Spatial Statistical Model for Landscape Genetics". *Genetics* 170(3) (2005), pp. 1261–1280.

[75] Wen-Yun Yang et al. "A model-based approach for analysis of spatial structure in genetic data". *Nat Genet* 44(6) (2012), pp. 725–731.

[76] John Michael Rañola, John Novembre, and Kenneth Lange. "Fast spatial ancestry via flexible allele frequency surfaces". *Bioinformatics* 30(20) (2014), pp. 2915–2922.

[77] Albert Jacquard. *Structures génétiques des populations*. Paris: Masson et Cie, 1970.

[78] Montgomery Slatkin. "Inbreeding coefficients and coalescence times". *Genetics Research* 58(2) (1991), pp. 167–175.

[79] Peter Mccullagh. *Structured covariance matrices in multivariate regression models*. 2006.

[80] Prem Gopalan et al. "Scaling probabilistic models of genetic variation to millions of humans". *Nat. Genet.* 48(12) (2016), pp. 1587–1590.

[81] Xiuwen Zheng and Bruce S. Weir. "Eigenanalysis of SNP data with an identity by descent interpretation". *Theoretical Population Biology*. New Developments in Relatedness and Relationship Estimation 107 (2016), pp. 65–76.

[82] Minsun Song, Wei Hao, and John D. Storey. "Testing for genetic associations in arbitrarily structured populations". *Nat Genet* 47(5) (2015), pp. 550–554.

[83] Jian Yang et al. "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1) (2011), pp. 76–82.

[84] Bowen Wang, Serge Sverdlov, and Elizabeth Thompson. "Efficient Estimation of Realized Kinship from SNP Genotypes". *Genetics* (2017), genetics.116.197004.

[85] Thomas Nagylaki. "Fixation Indices in Subdivided Populations". *Genetics* 148(3) (1998), pp. 1325–1332.

[86] Peter M. Visscher et al. "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings". *PLoS Genet.* 2(3) (2006), e41.

[87] Naoyuki Takahata and Masatoshi Nei. "FST and GST Statistics in the Finite Island Model". *Genetics* 107(3) (1984), pp. 501–504.

[88] Masatoshi Nei and Aravinda Chakravarti. "Drift variances of FSTand GST statistics obtained from a finite number of isolated populations". *Theoretical Population Biology* 11(3) (1977), pp. 307–325.

[89] Masatoshi Nei, Aravinda Chakravarti, and Yoshio Tateno. "Mean and variance of FST in a finite number of incompletely isolated populations". *Theoretical Population Biology* 11(3) (1977), pp. 291–306.

[90] Motoo Kimura and James F. Crow. "The Number of Alleles That Can Be Maintained in a Finite Population". *Genetics* 49(4) (1964), pp. 725–738.

[91] Takeo Maruyama. "Effective number of alleles in a subdivided population". *Theoretical Population Biology* 1(3) (1970), pp. 273–306.

[92] C. Clark Cockerham. "Analyses of Gene Frequencies". *Genetics* 74(4) (1973), pp. 679–700.

[93] M. Slatkin and L. Voelm. "FST in a hierarchical island model." *Genetics* 127(3) (1991), pp. 627–629.

[94] François Rousset. "Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance". *Genetics* 145(4) (1997), pp. 1219–1228.

[95] Rousset. "Genetic differentiation between individuals". *Journal of Evolutionary Biology* 13(1) (2000), pp. 58–62.

[96] Graham Coop et al. "Using Environmental Correlations to Identify Loci Underlying Local Adaptation". *Genetics* 185(4) (2010), pp. 1411–1423.

[97] B. D. H. Latter. "The island model of population differentiation: a general solution". *Genetics* 73(1) (1973), pp. 147–157.

[98] M. Slatkin. "A measure of population subdivision based on microsatellite allele frequencies." *Genetics* 139(1) (1995), pp. 457–462.

[99] L. Excoffier, P. E. Smouse, and J. M. Quattro. "Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data." *Genetics* 131(2) (1992), pp. 479–491.

[100] Yannis Michalakis and Laurent Excoffier. "A Generic Estimation of Population Subdivision Using Distances Between Alleles With Special Reference for Microsatellite Loci". *Genetics* 142(3) (1996), pp. 1061–1064.

[101] Masatoshi Nei. "F-statistics and analysis of gene diversity in subdivided populations". *Annals of Human Genetics* 41(2) (1977), pp. 225–233.

[102] M. Nei and R. K. Chesser. "Estimation of fixation indices and gene diversities". *Annals of Human Genetics* 47(3) (1983), pp. 253–259.

[103] Naoyuki Takahata. "Gene Identity and Genetic Differentiation of Populations in the Finite Island Model". *Genetics* 104(3) (1983), pp. 497–512.

[104] A. M. Bowcock et al. "Drift, admixture, and selection in human evolution: a study with DNA polymorphisms". *PNAS* 88(3) (1991), pp. 839–843.

[105] Rongwei Fu, Alan E. Gelfand, and Kent E. Holsinger. "Exact moment calculations for genetic models with migration, mutation, and drift". *Theoretical Population Biology*. Uses of DNA and genetic markers for forensics and population studies 63(3) (2003), pp. 231–243.

[106] Shuichi Kitada, Toshihide Kitakado, and Hirohisa Kishino. "Empirical Bayes Inference of Pairwise FST and Its Distribution in the Genome". *Genetics* 177(2) (2007), pp. 861–873.

[107] Michael D. Edge and Noah A. Rosenberg. "Upper bounds on in terms of the frequency of the most frequent allele and total homozygosity: The case of a specified number of alleles". *Theoretical Population Biology* 97 (2014), pp. 20–34.

[108] Bradley P. Carlin and Thomas A. Louis. "Bayes and empirical Bayes methods for data analysis". *Statistics and Computing* 7(2) (1997), pp. 153–154.

Supplementary Information: $F_{\mathbf{ST}}$ and kinship for arbitrary population structures I: Generalized definitions

Alejandro Ochoa[1,2] and John D. Storey[3,*]

[1]Duke Center for Statistical Genetics and Genomics, and [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

∗ Corresponding author: `jstorey@princeton.edu`

S1 Review of previous $F_{\mathbf{ST}}$ definitions

Here we review how F_{ST} measures the population structure of individuals, parametrizes genetic drift, and has been adapted to studying the genetic diversity at individual loci. Studies with different goals have demanded various definitions of F_{ST} as starting points, which have generated considerable confusion [22, 24, 48, 51, 53, 85]. Here we group these working definitions of F_{ST} into three classes and discuss their connection to our work. Note that all previous F_{ST} definitions are fundamentally about a subpopulation or a collection of disjoint subpopulations, and do not apply to individuals with arbitrary relatedness such as Hispanics (Section 2 and [54]) who have individual-specific admixture proportions (this admixture model is studied in Section 6 above and utilized to simulate data to benchmark generalized F_{ST} estimation in Section 6 of Part II).

S1.1 $F_{\mathbf{ST}}$ as a function of inbreeding coefficients

The original F_{ST} of Malécot and Wright is the mean inbreeding coefficient in a subpopulation S relative to an ancestral population T [5, 6], which corresponds to our f_S^T. If S is unstructured [5], then $f_j^T = f_S^T \forall j \in S$ in our notation and thus it can be given in terms of individual inbreeding coefficients by

$$F_{\mathrm{ST}} = \frac{1}{|S|} \sum_{j \in S} f_j^T. \tag{S1}$$

When S is structured [6] then three quantities are specified, which may be given in terms of individual coefficients by

$$F_{\mathrm{IT}} = \frac{1}{|S|} \sum_{j \in S} f_j^T, \quad F_{\mathrm{IS}} = \frac{1}{|S|} \sum_{j \in S} f_j^S, \quad F_{\mathrm{ST}} = \frac{F_{\mathrm{IT}} - F_{\mathrm{IS}}}{1 - F_{\mathrm{IS}}}. \tag{S2}$$

Note that when S is unstructured then Eq. (S2) reduces to Eq. (S1), since $f_j^S = 0 \forall j \in S$, so $F_{\mathrm{IT}} = 0$ and $F_{\mathrm{ST}} = F_{\mathrm{IT}}$. Additionally, Eq. (S2) holds under our generalized framework since

$\left(1 - f_j^T\right) = \left(1 - f_S^T\right)\left(1 - f_j^S\right)$, but the alternative form

$$F_{\mathrm{ST}} = \frac{1}{|S|}\sum_{j \in S} \frac{f_j^T - f_j^S}{1 - f_j^S}$$

is more directly comparable to our generalized definition of Eq. (3).

This original F_{ST} and the earlier inbreeding [67] and kinship coefficients [5] were all estimated from pedigrees rather than genetic markers as it is now more common. Thus, this F_{ST} measures only the relatedness of individuals in a subpopulation, it is independent of mutation rates or selection, and it is not defined by any particular genetic marker. Inbreeding coefficients and F_{ST} were estimated from a pedigree using the method of path coefficients [3]. Our generalized F_{ST}—defined in Eq. (3) using individual inbreeding coefficients—corresponds most closely to this original F_{ST} definition, with the important exception that we aim to estimate realized kinship coefficients rather than their expected values under the pedigree [86].

S1.2 F_{ST} as a model parameter of allele variance

Consider a biallelic locus i and some allele taken as a reference, which had an allele frequency p_i^T in the ancestral population T and which evolves to have an allele frequency p_i^S in a subpopulation S that derives from T such that the mean inbreeding of every individual in S relative to T is F_{ST}. T and S are implicitly panmictic populations, so that genotypes are in Hardy-Weinberg equilibrium and p_i^T and p_i^S suffice to describe their allele distributions. Wright found that for neutral loci i (without mutation and selection) the variance of the possible random values p_i^S that could result given fixed p_i^T and F_{ST} parameters is [31]

$$\mathrm{Var}\left(p_i^S \middle| T\right) = p_i^T\left(1 - p_i^T\right) F_{\mathrm{ST}}. \tag{S3}$$

Note again that this equation results from considering the effect that relatedness of individuals has on their allele frequencies at neutral loci, and that F_{ST} is thus shared across all such neutral loci.

Many subsequent works have taken Eq. (S3), restated as

$$F_{\mathrm{ST}} = \frac{\mathrm{Var}\left(p_i^S \middle| T\right)}{p_i^T\left(1 - p_i^T\right)},$$

to define F_{ST} [19, 22–24, 49, 51, 71, 87]. This alternative definition can lead to confusion for three reasons: it can be mistakenly interpreted as applying to all loci (including loci under mutation or selection); it suggests that every locus i has its own F_{ST}; and it depends on the unknowns $\mathrm{Var}\left(p_i^S \middle| T\right)$ and p_i^T that have been interpreted in various ways [22, 24, 48, 49, 51–53, 88, 89]. We stress that Wright and Malécot originally defined F_{ST} from inbreeding coefficients and Eq. (S3) was derived as a consequence of the relatedness of individuals (as captured by F_{ST}) and applies only to neutral loci [5, 31].

Complicating matters, in developing the effect of F_{ST} on allele distributions, both Malécot and Wright extended F_{ST} to incorporate the effect of mutation into Eq. (S3), resulting in formulas for the F_{ST} in a population at equilibrium, such as

$$F_{\mathrm{ST}} \approx \frac{1}{4N\mu + 1},$$

for a population with N individuals at all times, where μ is the sum of migration (proportion of individuals N per generation) and mutation rates (proportion of mutations per locus per generation), and F_{ST} above is the approximate value approached in infinite time [5, 6]. Both authors note that migration reduces F_{ST} in the inbreeding definition, and mutation has an identical mathematical effect on reducing the variance of p_i^S, thus mutation reduces the *effective* F_{ST} by reducing the probability of allele fixation [5, 6]. In contrast, the inbreeding F_{ST} approaches 1 with infinite time in a finite and isolated population [69], regardless of mutation. Since locus mutation does not alter inbreeding values, this extended F_{ST} that captures mutation is no longer compatible with the inbreeding F_{ST}. Many later works with greater focus on the evolution of allele frequencies than on relatedness adopted the F_{ST} definition that incorporates mutation [16, 17, 51, 78, 90–94]. Frameworks that assume neutral loci only—and thus are compatible with the inbreeding F_{ST}—include our work, method-of-moments F_{ST} estimators [17, 22–24, 46, 47, 95] and Normal [23, 57, 96] and Bayesian likelihood models based on the Beta (for biallelic loci) or Dirichlet (multiallelic) distributions [7, 20, 36, 61] for the subpopulation allele frequencies p_i^S. Some authors model F_{ST} and mutation as separate effects [97, 98].

In the coalescent framework, in the limit of small mutation rates, F_{ST} was shown to equal

$$F_{\mathrm{ST}} = \frac{\bar{t}_T - \bar{t}_S}{\bar{t}_T},$$

where $\bar{t}_T$ and $\bar{t}_S$ are average coalescence times for alleles within the populations T and S, respectively [78]. This connection to coalescent times led to the R_{ST} statistic that remarkably estimates F_{ST} from microsatellites while excluding the effect of the relatively high mutation rate of these variants under some assumptions [98]. The Weir-Cockerham F_{ST} estimator and R_{ST} are special cases of ϕ_{ST} in the AMOVA framework [99, 100].

S1.3 F_{ST} as a data-dependent statistic that measures variance at a locus

Locus-specific F_{ST} estimates are often employed to identify loci under selection [14–21]. In this setting, F_{ST} is often defined by the following sample estimate of Eq. (S3) per locus i,

$$\hat{F}_{\mathrm{ST},i}^{\mathrm{sample}} = \frac{\hat{\sigma}_i^2}{\hat{p}_i^T\left(1 - \hat{p}_i^T\right)}, \tag{S4}$$

where $p_i^{S_u}$ is the reference allele frequency in each subpopulation S_u, p_i^T is estimated by the sample mean over the K subpopulations $\hat{p}_i^T = \frac{1}{K}\sum_{u=1}^{K} p_i^{S_u}$, and $\mathrm{Var}\left(p_i^S \middle| T\right)$ is estimated by the sample

variance $\hat{\sigma}_i^2 = \frac{1}{K}\sum_{u=1}^{K}\left(p_i^{S_u} - \hat{p}_i^T\right)^2$ [48, 51–53, 85, 88, 89, 101–106]. Unlike the random variable p_i^S in the $F_{\rm ST}$ definition of Eq. (S3), studies of Eq. (S4) usually treat $p_i^{S_u}$ as fixed parameters. Thus, although $\hat{F}_{{\rm ST},i}^{\rm sample}$ shares many of the properties of $F_{\rm ST}$, $\hat{F}_{{\rm ST},i}^{\rm sample}$ is a biased estimator of the $F_{\rm ST}$ from Eq. (S3) [22, 71], so these definitions are not compatible. Nevertheless, since $\hat{F}_{{\rm ST},i}^{\rm sample}$ is effective for studying the evolution of individual loci, it has spawned its own field of research, starting with Nei's $G_{\rm ST}$ that generalizes $\hat{F}_{{\rm ST},i}^{\rm sample}$ to multiple alleles [48] and is often treated as $F_{\rm ST}$ [52, 53, 78, 85, 88, 89, 102, 103, 106], related single-locus $F_{\rm ST}$ estimators based on method-of-moments [16–19] or Bayesian models [20, 21, 96, 106], and alternative quantities such as $G'_{\rm ST}$ [49] and D [50] ($G_{\rm ST}$ approximates $F_{\rm ST}$ better than $G'_{\rm ST}$ and D, especially under a low mutation rate [51]). Note that although $F_{\rm ST}$ was previously studied for biallelic loci only [5, 6, 31, 48], there are more recent $F_{\rm ST}$ models that generalize Eq. (S3) for neutral multiallelic loci [7, 22, 23] analogous to how the $G_{\rm ST}$ statistic generalizes Eq. (S4). Locus-specific $F_{\rm ST}$ estimates present unique challenges since their sampling distribution depends on demography and heterozygosity or the maximum allele frequency at the locus [16–20, 49, 53, 104, 107]. The focus of our work is to generalize and accurately estimate the genome-wide $F_{\rm ST}$ in individuals with arbitrary relatedness, and does not presently concern locus-specific $F_{\rm ST}$ estimation or the identification of loci under selection.

S2 Derivation of kinship and $F_{\rm ST}$ in terms of mean coalescence times

We shall consider the probability of identity by descent (IBD) in a random process that admits mutations along the coalescent tree. Interestingly, the limit as the mutation rate goes to zero results in non-trivial connections between the IBD coefficients and coalescence times. Our proof closely mirrors that of [78].

Let μ be the mutation rate, in units of mutations per base per generation, which is assumed to be a constant for all branches of the tree. Let h_1 and h_2 denote two haploid DNA sequences (we shall convert to diploid individuals in the end). Let $P_{h_1h_2}(t)$ be the probability that h_1 and h_2 coalesce in generation t. By definition, the sum of these probabilities across all coalescence times ($t \geq 1$) equals one:

$$\sum_{t=1}^{\infty} P_{h_1h_2}(t) = 1.$$

The overall probability that a given random locus at both sequences is IBD is the expectation of $(1-\mu)^{2t}$—the probability that a mutation has not occured by generation t at this locus for both sequences h_1 and h_2:

$$g_{h_1h_2}(\mu) = {\rm E}_t\left[(1-\mu)^{2t}\middle|h_1,h_2\right] = \sum_{t=1}^{\infty}(1-\mu)^{2t}P_{h_1h_2}(t).$$

Note that $g_{h_1h_2}(0) = 1$ and

$$g'_{h_1h_2}(\mu) = -\sum_{t=1}^{\infty} 2t(1-\mu)^{2t-1} P_{h_1h_2}(t), \quad \text{so}$$

$$g'_{h_1h_2}(0) = -\sum_{t=1}^{\infty} 2t P_{h_1h_2}(t) = -2\,\mathrm{E}_t\left[t|h_1,h_2\right] = -2\bar{t}_{h_1h_2},$$

where $\bar{t}_{h_1h_2}$ is the mean coalescence time of sequences h_1 and h_2. To proceed, consider the equivalent quantity for the two most distant sequences in the sample, which are taken as being drawn independently from the ancestral population T:

$$g_T(\mu) = \sum_{t=1}^{\infty} (1-\mu)^{2t} P_T(t).$$

The IBD coefficient of interest, $g^T_{h_1h_2}(\mu)$, is a relative probability related to $g_{h_1h_2}(\mu)$ and $g_T(\mu)$ in the same manner as F_{ST}, F_{IT} and F_{IS}, namely

$$\left(1 - g_{h_1h_2}(\mu)\right) = \left(1 - g^T_{h_1h_2}(\mu)\right)\left(1 - g_T(\mu)\right).$$

Note that solving for $g^T_{h_1h_2}(0)$ above gives an undefined value $(0/0)$, since $g_{h_1h_2}(0) = g_T(0) = 1$. Nevertheless, solving for $g^T_{h_1h_2}(\mu)$ first (for $\mu \neq 0$) and taking the limit as the mutation rate goes to zero (using L'Hôpital's rule), we obtain the IBD probability of interest, for h_1 and h_2 relative to T:

$$\begin{aligned} f^T_{h_1h_2} &= \lim_{\mu\to 0} g^T_{h_1h_2}(\mu) \\ &= \lim_{\mu\to 0} \frac{g_{h_1h_2}(\mu) - g_T(\mu)}{1 - g_T(\mu)} \\ &= \frac{g'_{h_1h_2}(0) - g'_T(0)}{-g'_T(0)} \\ &= \frac{\bar{t}_T - \bar{t}_{h_1h_2}}{\bar{t}_T}. \end{aligned}$$

The coefficients of interest are special cases of the last expression, as follows. The inbreeding coefficient is

$$\begin{aligned} f^T_j &= \frac{\bar{t}_T - \bar{t}_j}{\bar{t}_T}, \\ \bar{t}_j &= \bar{t}_{j_1j_2}, \end{aligned}$$

where j_1 and j_2 are the two haplotypes of individual j (the maternal and paternal alleles). Similarly, the kinship coefficient is an average of haplotype comparisons across individuals,

$$\begin{aligned} \varphi^T_{jk} &= \frac{\bar{t}_T - \bar{t}_{jk}}{\bar{t}_T}, \\ \bar{t}_{jk} &= \frac{1}{4}\left(\bar{t}_{j_1k_1} + \bar{t}_{j_1k_2} + \bar{t}_{j_2k_1} + \bar{t}_{j_2k_2}\right), \end{aligned}$$

where j_1 and j_2 are the two haplotypes of individual j, and k_1 and k_2 are the two haplotypes of individual k.

S3 Empirical Bayes estimation of subpopulation allele frequencies for map

The allele frequencies shown in the map of Fig. 1B are estimated from genotypes using Empirical Bayes with a Beta prior [108], as follows. Let x_{ij} be the number of reference alleles at locus i and subpopulation j, and n_{ij} be the total number of alleles. We model the desired subpopulation allele frequencies π_{ij} as drawn independently from a Beta prior:

$$\pi_{ij} \sim \text{Beta}(\alpha_i, \beta_i),$$
$$x_{ij}|\pi_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij}).$$

The marginal distribution of x_{ij} is the Beta-Binomial. The posterior estimate of π_{ij} that was displayed in Fig. 1B is

$$\hat{\pi}_{ij} = \frac{x_{ij} + \alpha_i}{n_{ij} + \alpha_i + \beta_i},$$

which compared to the sample estimate $\frac{x_{ij}}{n_{ij}}$ is "shrunk" toward the prior mean $p_i = \frac{\alpha_i}{\alpha_i + \beta_i}$ depending on sample size ($n_{ij} \gg \alpha_i + \beta_i$ have $\hat{\pi}_{ij}$ close to $\frac{x_{ij}}{n_{ij}}$, while $n_{ij} \ll \alpha_i + \beta_i$ have $\hat{\pi}_{ij}$ closer to p_i).

Instead of choosing α_i, β_i *a priori*, in empirical Bayes estimation α_i, β_i are the values that maximize the log-likelihood of the data,

$$\sum_j \log \ell(\alpha_i, \beta_i; x_{ij}, n_{ij}),$$

where ℓ is the Beta-Binomial likelihood function.

The Human Origins dataset was processed as described in [54], and additionally filtered to consider only loci with a minor allele frequency $\geq 10\%$ (362,437 loci) in identifying the locus with the median per-locus Weir-Cockerham F_{ST} estimate (using the $K = 244$ sub-subpopulations to partition individuals). For the locus rs2650044 displayed on Fig. 1B we estimated $\alpha_i \approx 1.83$ and $\beta_i \approx 8.34$.

S4 Proof that expected heterozygosity is indepenent of T

Here we show that H_{ij} has the same form conditioned on some ancestral population S as it does for any other choice T ancestral to S. Conditional on T, all of p_i^T, f_j^T and f_j^S are constant parameters, but p_i^S is a random allele frequency that drifted from the more ancestral p_i^T frequency, so p_i^S must be marginalized. Therefore, it suffices to prove that

$$\text{E}\left[p_i^S\left(1 - p_i^S\right)\middle|T\right]\left(1 - f_j^S\right) = p_i^T\left(1 - p_i^T\right)\left(1 - f_j^T\right).$$

We assume that p_i^S satisfies the coancestry model of Eqs. (9) and (10), yielding:

$$\text{E}\left[p_i^S\middle|T\right] = p_i^T,$$
$$\text{Var}\left(p_i^S\middle|T\right) = p_i^T\left(1 - p_i^T\right)f_S^T,$$

where f_S^T is the inbreeding coefficient of population S relative to T (see Section 3.1) and satisfies the IBD shift identity in Eq. (5):

$$\left(1 - f_j^T\right) = \left(1 - f_j^S\right)\left(1 - f_S^T\right).$$

The desired conclusion follows:

$$\begin{aligned}\mathrm{E}\left[p_i^S\left(1 - p_i^S\right)\middle|T\right] &= \mathrm{E}\left[p_i^S\middle|T\right] - \left(\mathrm{Var}\left(p_i^S\middle|T\right) + \left(\mathrm{E}\left[p_i^S\middle|T\right]\right)^2\right) \\ &= p_i^T\left(1 - p_i^T\right)\left(1 - f_S^T\right) = p_i^T\left(1 - p_i^T\right)\frac{\left(1 - f_j^T\right)}{\left(1 - f_j^S\right)}.\end{aligned}$$